Appunti per la definizione di un set di metadati gestionali-amministrativi e strutturali per le risorse digitali. – Versione 0 [1]

[1] Versione 0 del 2002-05-03 / preparata da A. Scolari, M. Pepe, M. Messina, C. Leombroni, G. Cirocchi, G. Bergamin per il *Gruppo di studio sugli standard e le applicazioni di metadati nei beni culturali* promosso dall'ICCU. Per gli obiettivi del gruppo vedi http://www.iccu.sbn.it/grupmeta.htm. Nel documento si farà riferimento a questa tipologia di metatati con l'acronimo **MAG.**

That has got to rank as one of the most overused and misused words in the semi-technical vocabulary of the early 21st century.
The word "metadata" means "data that describes or characterizes other data". Thus, nearly anything can be considered to be "metadata" under some circumstance. Conversely, almost everything that can be called "metadata" can also be called "data"
When you are processing a file of catalog records, they are data. When you are considering them in their role as descriptors for books, they are metadata.
(Michael McClennen, 17-4-2002, nella lista PERL4LIB)

1. Contesto

Questo documento costituisce la seconda versione (denominata *versione 0*) della proposta presentata come *bozza preliminare* nei primi mesi del 2001 e reperibile in http://www.iccu.sbn.it/metaAG1.pdf.

L'impianto del documento originale è stato interamente rivisto per prendere in conto in maniera particolare:

- la proposta della Library of Congress di METS (Metadata Encondig & Transmission Standard)[2];
- i lavori del gruppo congiunto OCLC/RLG sui medatati per la conservazione delle risorse digitali[3]

2. Scopo del lavoro

Produrre uno Schema XML basato su METS e sull'esperienza acquisita nell'ambito del progetto ARSBNI. Scopo dello Schema è quello di dare delle specifiche formali per la fase di raccolta e di archiviazione dei metadati e dei dati digitali.

[2] http://www.loc.gov/standards/mets/<<The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.>>
[3] http://www.rlg.org/longterm/<<Calling on the deep strengths of participants in RLG's preservation program, PRESERV, work under this initiative will be coordinated with partner organizations such as the Digital Library Federation (DLF), the National Library of Australia, the Council on Library and Information Resources (CLIR), and OCLC, to eliminate redundant effort and ensure maximum impact>>

Lo Schema è composto dalle seguenti sezioni:

- **gen** informazioni generali sul progetto e sul tipo di digitalizzazione
- **bib** metadati descrittivi sull'oggetto digitalizzato
- **img** metadati specifici relativi alle immagini fisse
- **ocr** metadati specifici relativi al riconoscimento ottico del testo

E' previsto l'inserimento nello Schema delle sezioni relative al suono ed alle immagini in movimento ed un'evoluzione per il trattamento dei metadati nel caso di oggetti digitali nativi.

Lo Schema è aperto, ma il sottogruppo MAG è partito dall'assunzione che il *core* dei MAG sia quello indicato nella versione preliminare di questo documento

Con questo Schema XML all'interno di un progetto di digitalizzazione ogni postazione di lavoro è in grado di produrre per ogni oggetto digitalizzato un file guida "standard" ovvero conforme allo Schema XML che:
- raccoglie tutte le informazioni sull'oggetto digitalizzato (metadati);
- contiene la mappa di tutti i file generati contestualmente dalla digitalizzazione e relativi all'oggetto digitalizzato

Lo Schema XML non fa naturalmente "tutto". Si ritiene tuttavia che qualsiasi sistema professionale di digitalizzazione sia in grado in poco tempo di attrezzarsi per produrre un "output standard": l'insieme di un file guida e di uno o più file immagine/ocr ecc.

In altre parole il sottogruppo MAG non ritiene utile perseguire la strada del software "unico per tutti e che fa tutto", ma ritiene essenziale che in questo campo i produttori di servizi informatici siano messi in concorrenza a partire da standard aperti, documentati e liberamente disponibili.

3. Lavori da fare

Grazie alla standardizzazione operata in fase di acquisizione (indipendente da un determinato hardware e da un determinato software) è possibile popolare gli archivi - "i magazzini del digitale" - in maniera standardizzata .

Il sottogruppo MAG produrrà entro la fine di maggio un documento di raccomandazione sull'architettura *logica* dei "magazzini del digitale" pensati per "durare nel lungo periodo".

I programmi di fruizione potranno essere così costruiti a partire da una infrastruttura standard: occorre evitare che a tipi di fruizione diversa corrispondano archiviazione di tipo diverso (p. es. le immagini scelte di un libro pubblicate su web in una mostra virtuale e il libro completo visualizzabile dall'OPAC dovrebbero attingere allo stesso archivio).

4. Metadati e progetti di digitalizzazione

Nei progetti di digitalizzazione e nelle attività di gestione degli archivi di oggetti digitali i metadati rivestono un'importanza crescente, tanto da venire considerati parte costituente della definizione di Oggetto Digitale[4]. La realizzazione di una biblioteca digitale presuppone quindi un'accurata

[4] Si veda ad es. la definizione di Oggetto Digitale in California Digital Library, *Digital Object Standard: Metadata, Content and Encoding*, May 18, 2001: "Un oggetto digitale è definito … come un qualcosa (es.

definizione dei metadati da associare agli oggetti che la compongono. Le istituzioni della memoria hanno sempre avuto a che fare con i metadati, la cui definizione più semplice è quella di "informazioni su altre informazioni". In una scheda di catalogo, la descrizione bibliografica è un metadato di natura descrittiva, mentre il numero d'inventario e la segnatura sono metadati di natura amministrativa e gestionale. Ha scritto Day che "i metadati sono comunemente intesi come un'amplificazione delle tradizionali pratiche di catalogazione bibliografica in un ambiente elettronico"[5].
Vi sono diverse proposte di tassonomia dei metadati. Ai fini di questo documento i metadati possono essere distinti in due macrocategorie funzionali:

- DESCRITTIVI: servono per l'identificazione ed il recupero degli oggetti digitali; sono costituiti da descrizioni normalizzate dei documenti fonte (o dei documenti nati in formato digitale), risiedono generalmente nelle basi dati dei sistemi di *Information Retrieval* all'esterno dell'archivio digitale, e sono collegati a quest'ultimo tramite appositi *link*.

- AMMINISTRATIVI e GESTIONALI: evidenziano le modalità di archiviazione e manutenzione degli oggetti digitali nel sistema di gestione dell'archivio digitale, e sono necessari per una corretta esecuzione delle relative attività.

Distinzioni più articolate fra tipologie diverse di metadati possono essere fatte sulla base delle loro funzioni specifiche, ad esempio mappandone le tipologie sui diversi ruoli dei metadati all'interno del modello logico-funzionale OAIS dell'archivio digitale[6].

Nel mondo digitale, data la labilità dell'informazione elettronica, i metadati amministrativi e gestionali assumono un'importanza preponderante ai fini della conservazione permanente degli oggetti digitali. Essi possono "documentare i processi tecnici associati alla conservazione permanente, fornire informazioni sulle condizioni ed i diritti di accesso agli oggetti digitali, certificare l'autenticità e l'integrità del contenuto, documentare la catena di custodia degli oggetti, identificarli in maniera univoca"[7].

Nel lavoro di definizione di un primo set standard di metadati amministrativi e gestionali si è tenuto conto delle principali iniziative e dei più significativi progetti realizzati in ambito internazionale. Sulla base di tali esperienze si ritiene di mettere in evidenza quanto segue:

- più che per particolari tipologie di documenti fonte (es. periodici, musica a stampa o manoscritta, carte geografiche etc.), il set di metadati di base viene definito per tipologie di oggetti digitali; una prima distinzione può riguardare:
  - immagini statiche,
  - testi prodotti con tecnologia ocr,
  - suono,
  - immagini in movimento e oggetti multimediali,
  - *born digital*;

---

un'immagine, una registrazione audio, un documento testuale) che è stato codificato in modo digitale e integrato con metadati tali da supportarne l'individuazione, l'uso e l'immagazzinamento".
http://www.cdlib.org/about/publications/CDLObjectStd-2001.pdf

[5] M. Day, *Issues and Approaches to Preservation Metadata*. In: Proceedings from the Joint RLG and NPO Preservation Conference, September 1998, http://www.rlg.org/preserv/joint/day.html

[6] Per una descrizione del modello OAIS si veda il par. 2.

[7] Liberamente tradotto da: OCLC/RLG Working Group on Preservation Metadata, *Preservation Metadata for Digital Objects: a Review of the State of the Art, a White Paper*, January 2001, p. 2. http://www.oclc.org/research/pmwg/presmeta_wp.pdf

- i metadati ed il loro sistema di gestione devono essere completamente indipendenti da specifiche piattaforme HW e SW, al fine di favorirne un impiego generalizzato;
- devono invece essere coerenti con le funzioni previste nel modello logico-funzionale standard dell'archivio degli oggetti digitali cui si fa riferimento (es. le funzioni di Immissione, Archiviazione, Gestione, Accesso, Amministrazione, Pianificazione della conservazione, nel modello OAIS).

5. Il modello OAIS

A partire dal *Rapporto* del 1996 della Task Force on Archiving of Digital Information[8], che per primo ha inquadrato la complessa problematica della conservazione permanente delle risorse digitali, nel suo duplice aspetto di mantenimento degli insiemi di bit che le compongono e di continuità di accesso al loro contenuto, si sono avuti interessanti sviluppi.

Importanti iniziative internazionali hanno infatti lavorato sulle indicazioni contenute nel Rapporto, in particolare sulla necessità di costituire una *infrastruttura profonda* in grado di sostenere un sistema distribuito di archivi digitali.

Il modello logico-funzionale denominato OAIS è parso adeguato per rispondere a questa esigenza, ed è stato adottato come riferimento da diverse esperienze in Europa, negli Stati Uniti e in Australia.[9]

Un ulteriore sviluppo delle raccomandazioni della Task Force, che incorpora il lavoro finora svolto sul modello OAIS, è la collaborazione tra RLG e OCLC per la specificazione degli attributi di un archivio digitale certificato.[10]

L'acronimo OAIS è traducibile con **Sistema Informativo Aperto per l'Archiviazione**: si tratta di uno standard ISO in fase di elaborazione attraverso un processo, per l'appunto, "aperto" ai contributi di enti e istituzioni diverse.[11]

Sviluppato in origine dalla comunità della ricerca spaziale, il modello OAIS è adeguato alle finalità della conservazione permanente anche per altri tipi di comunità; pur essendo fortemente orientato al trattamento dei documenti elettronici, esso non specifica nessun tipo di implementazione ed è utilizzabile per qualsiasi tipo di archivio - digitale o analogico: può essere applicato indifferentemente a oggetti digitali nativi, a prodotti di attività di digitalizzazione (quali file di immagini), e persino a oggetti fisici.
L'OAIS si autodefinisce *"un'organizzazione di soggetti e sistemi che hanno accettato la responsabilità della conservazione dell'informazione e del mantenerla disponibile per una comunità*

[8] John Garrett, Donald Waters, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access and RLG, Washington, D.C., 1996 in www.rlg.org/ArchTF/index.html

[9] Tra di esse può essere citata in primo luogo NEDLIB - Networked European Deposit Library, che ha formulato un modello di deposito delle pubblicazioni elettroniche, poi il progetto CEDARS (CURL Exemplars in Digital Archives), promosso dalle Università di Cambridge, Oxford e Leeds, che ha sviluppato uno schema di metadati per la conservazione basato sul modello OAIS, e PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) della National Library of Australia, che mira alla conservazione di siti Web selezionati. Ad esso si e' ispirata implicitamente la Library of Congress per la definizione del proprio set di metadati, oltre che la BNF per il progetto ARSBNI.

[10] *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources* / an RLG-OCLC Report. - Draft for Public Comment, August 2001 in http://www.rlg.org/longterm/attributes01.pdf

[11] *Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-R-2.* Red Book (Draft Standard). Issue 2. June 2001 in http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

determinata"; come tale, si propone di fornire una solida base per una ulteriore standardizzazione nel contesto dell'archiviazione digitale e di promuovere una maggiore consapevolezza, da parte dei venditori di tecnologie, delle esigenze della conservazione a lungo termine dei documenti elettronici.

Riconoscendo che le collezioni digitali sono caratterizzate da una natura fortemente distribuita, e che vi è necessità, al tempo stesso, di sviluppare localmente politiche e procedure efficaci di gestione e conservazione di tali risorse, esso delinea un modello di archiviazione distribuita, ma rispondente ad un ben individuato modello logico.

Il modello non esprime alcuna preferenza riguardo alle strategie di conservazione dei dati (p. es. migrazione o emulazione), ma individua termini e concetti rilevanti per l'archiviazione di documenti digitali, identifica le componenti e i processi chiave comuni alla maggior parte delle attività di conservazione digitale, e propone un **modello logico di riferimento** per gli oggetti digitali e i metadati loro associati.

Dall'insieme di queste caratteristiche è scaturita una rapida affermazione dell'OAIS: la tematica della *responsabilità* della conservazione, fortemente accentuata, lo ha raccomandato come standard adeguato alla creazione di un archivio di oggetti digitali ad importanti esperienze nel mondo delle biblioteche, degli archivi, delle istituzioni museali e di tutte quelle realtà che si pongono come obiettivo la conservazione a lungo termine dei documenti loro affidati.

Viene di seguito delineata una rapida sintesi dell'ambiente, delle componenti funzionali e della tassonomia degli oggetti informativi presentati dal modello.

5.1 Ambiente e componenti funzionali

L'**ambiente** OAIS è formato dall'interazione di quattro **entità**, definite come:
- i produttori (che sulla base di accordi formalizzati per l'immissione dei dati nell'archivio forniscono i dati in base alle componenti logiche e al modello di rappresentazione OAIS)
- i consumatori (le persone, o sistemi clienti, che interagiscono con i servizi OAIS per reperire e acquisire l'informazione conservata di proprio interesse).
- il management (esterno all'archivio, si occupa delle politiche dell'archivio: cosa archiviare, come trovare i fondi ecc.)
- l'archivio stesso (inteso come organizzazione che si propone di conservare l'informazione per consentirne l'accesso e l'uso ad una Comunità Designata)

Le **componenti funzionali** dell'OAIS sono definite come:
- Immissione (Ingest): in questa fase si riceve l'informazione dai produttori e la si prepara per l'archiviazione
- Archiviazione dei dati (Archival Storage): tratta l'archiviazione, la manutenzione e la gestione della informazione archiviata
- Gestione dei dati (Data Management): coordina i metadati *descrittivi* relativi *sia* all'informazione archiviata *che ai dati amministrativi interni all'archivio*
- Accesso (Access): è la funzione che aiuta i consumatori a identificare e ottenere informazione dall'archivio
- Amministrazione dell'archivio (Administration): ha in carico le operazioni giornaliere di mantenimento dell'archivio.
- Pianificazione della conservazione (Preservation Planning) che specifica e rende "visibili" ulteriormente le attività di conservazione effettuate, a prescindere dalla strategia adottata.

5.2 Classificazione degli Oggetti Informativi

Oltre alle componenti funzionali, l'OAIS fornisce anche un modello di strutturazione dei dati adeguato a rappresentare l'informazione digitale da un punto di vista orientato alla conservazione.

Fondamentale è la definizione di **Oggetto Informativo** (Information Object), di qualsiasi tipo esso sia, come composto da due elementi:

- **Dati**
- **Informazione sulla Rappresentazione (Metadati),** necessaria per conferire significato ed interpretabilità ai dati.

La conservazione nel tempo dei due elementi è un requisito fondamentale nel funzionamento del modello: in un ambiente digitale, i **Dati** (uno o più insiemi di bit) devono necessariamente restare collegati ad una **Informazione sulla Rappresentazione** (Representation Information) che contenga tutto ciò che è necessario per rendere comprensibili quei bit, sia dal punto di vista *strutturale* (specificazione del formato, descrizione del s/w di accesso, etc.), che *semantico* (p. es. in quale lingua è un testo in caratteri ASCII). [12]

Ogni scambio di informazione, da e per l'archivio, e all'interno di OAIS, avviene attraverso l'utilizzazione di **Pacchetti di Informazione** (Information Packages), che sono dei contenitori concettuali - di tipo logico non fisico - di dati.

Gli **Oggetti Informativi** possono comporre tre tipi di **Pacchetti di Informazione**, utilizzati a seconda che il flusso di informazioni avvenga dal Produttore all'OAIS, al suo interno, o dall'archivio stesso verso il Consumatore:

- **SIP** - (*Submission Information Package)* - Pacchetto di Informazioni per l'Immissione, utilizzato nella fase di immissione/acquisizione dei dati (alla fase di raccolta dei dati il gruppo MAG ha finora rivolto la propria attenzione, sulla scia dell'esperienza condotta in BNCF)
- **AIP** - (*Archival Information Package*) - Pacchetto di Informazioni per l'Archiviazione, destinato alla conservazione a lungo termine.
- **DIP** - (*Dissemination Information Package*) - Pacchetto di Informazioni per la Distribuzione, trasferito dall'OAIS all'utente in base ad una richiesta di accesso.

Ogni **Pacchetto di Informazione** è sempre costituito dall'aggregazione di quattro classi di **Oggetti Informativi:**

1. Informazione sul Contenuto / **Content Information**: l'oggetto primario (i dati) destinato ad essere conservato dall'archivio: consiste degli oggetti digitali primari e di tutte le Informazioni

[12] Ma anche per un documento fisico vale lo stesso modello di strutturazione; per es., l'informazione contenuta in un libro in italiano, espressa da caratteri a stampa (i dati), in combinazione con una conoscenza della lingua del testo (la Conoscenza di Base) si converte in informazione. Se la lingua italiana non rientra tra le conoscenze del ricevente, il testo (i dati) dovrà essere accompagnato da un dizionario italiano e da informazioni grammaticali espresse in una forma compresa nella Conoscenza di Base di chi riceve l'informazione: questi costituiranno **l'Informazione sulla Rappresentazione**. Si può dire che "*i Dati, in unione alla loro Informazione sulla Rappresentazione, veicolano significato*".

sulla Rappresentazione necessarie a trasformare i dati in informazioni significative (p.es. un documento in formato PDF (i dati), più la documentazione del formato PDF (la R.I.)[13]

2. Informazione Descrittiva per la Conservazione / **Preservation Description Information (PDI)**. E' il set di informazioni (metadati) necessarie a conservare adeguatamente, per un periodo di tempo indefinito, il Contenuto cui sono associate; si focalizza sulla descrizione degli stati passati e presenti del Contenuto, assicura che sia identificato univocamente, e che non sia alterato senza che ciò venga registrato.[14]

Il Contenuto e l'Informazione Descrittiva per la Conservazione (PDI) sono viste come strettamente collegate e identificabili per mezzo della:

3. Informazione sulla Composizione del Pacchetto di Informazione ovvero Informazione di Impacchettamento (**Packaging Information**): metadati utili a reperire 1) e 2) ovvero a evidenziare come sono collegati i componenti di un Pacchetto di Informazioni in un'entità identificabile su uno specifico supporto (p. es. in che disco, in che directory ecc.)

Il "pacchetto di informazione" che ne risulta è ricercabile attraverso la

4. Informazione Descrittiva (**Descriptive Information**): Metadati finalizzati alla ricerca e al recupero dell'informazione, p. es.: Manzoni, Alessandro - Autore principale..

Seguendo quindi il modello OAIS i metadati potrebbero essere distinti come segue:

- Metadati collegati alla Informazione Descrittiva (Description Information) [descrittivi][15]
- Metadati collegati alla Rappresentazione dell'Informazione (Representation Information) [tecnici]
- Metadati collegati alla struttura dell'oggetto informativo (Packaging Information) [strutturali]
- Metadati collegati alla Conservazione (Preservation Description Information) [per la conservazione]

Altri metadati, quali quelli relativi alla gestione dei diritti d'accesso, benchè rilevanti per gli obiettivi del gruppo MAG, non sono presi direttamente in considerazione dal modello OAIS.

---

[13] Su questo punto è importante il lavoro di approfondimento svolto dal gruppo di lavoro congiunto OCLC-RLG, vedi OCLC-RLG Working Group on Preservation Metadata, *A Recommendation on Content Information : a Report*, october 2001, in http://www.oclc.org/research/pmwg/

[14] l'OAIS individua quattro categorie all'interno della PDI (**Informazione Descrittiva per la Conservazione**); esse sono:

- **Identificazione** (Reference Information): enumera e descrive gli identificatori assegnati al Contenuto (p. es. l'URN)
- **Contesto** (Context Information): documenta le relazioni del Contenuto con il suo ambiente (perché è stato creato, in che relazione è con altri Contenuti) (p.es. altri O.I. nella stessa collezione, manifestazioni precedenti dello stesso Oggetto)
- **Provenienza** (Provenance Information): documenta la storia del Contenuto e i cambiamenti da esso subiti, oltre che la catena di custodia (p. es. il formato originale dei dati, quale specifico processo ha permesso la creazione /trasformazione di un determinato oggetto informativo, chi è stato responsabile di quel processo)
- **Autenticazione** (Fixity Information): documenta i meccanismi di autenticazione destinati ad assicurare l'integrità del Contenuto (p. es. l'impronta digitale calcolata con l'algoritmo MD5).

[15] Tra parentesi quadra il nome che viene talvolta usato.